CMIP6 "Impact" on Scientific Community

Sergey Nikonov[1], V.Balaji[1], Erik Mason[2], Aparna Radhakrishnan[2], Nalanda Sharadjaya[3], Hans Vahlenkamp[4]

[1] Princeton University, NJ

[2] Engility, NJ

[3] Stuyvesant High School, New York

[4] UCAR, CO

6th Annual ESGF F2F Conference.
Dec 6-9, 2016. Washington, DC.

Outline

- Comparison of 3 IPCCs: AR4, AR5, AR6
- Resources spent for AR5: Data Producers vs Data Consumers
- Usage of AR5 – download, analysis, scientific outcome
- Expected human costs for CMIP6 on Data Producers side.
- Potential efforts for CMIP6 data scientific acquisition.

CMIP3→ CMIP5 → CMIP6 Evolution (or Revolution)

CMIP Experience

- This is my 3rd CMIP in my life and it's getting more and more exciting.
- IPCC AR 4 (CMIP3) was the challenge for GFDL IT capabilities – computational and bandwidth resources. We had bottleneck in CMORizing and transferring data from archive to Data Portal/PCMDI. FedEx data transfer to PCMDI happened faster than ftp. The volume of GFDL data was just 12 TB.
- CMIP5 was much better from IT point of view. We've got Curator system for that. Main challenge happened in scientific manmade QC. It was a essential burden for GFDL scientists.
- The team was about 10 scientist and goal to make QC of ~200 TB of data diversified into 600 variables and saved into 1 million files.

CMIP5 was a Challenge

- Number of Experiments: 40
- Data diversification: 20 CMIP tables
- Number of Variables: 1000
- Number of Years: 5500
- Total Amount of Data Generated: 1.7 PB
- GFDL Amount: 180 TB

AR4/AR5 GFDL Download Growth and AR6 Projection

Project	Amount, TB Download/ Saved / Ratio	Files Download/Saved / Ratio	Hosts requesting	Averaged Bandwidth
IPCC AR4	150 / 12 / 13	5.2e+5 / 2.8e+4 / 20	4000	10 Mbit/sec
IPCC AR5	1300 / 180 / 7	8e+6 / 8.5e+5 / 10	9000	70 Mbit/sec
Growth AR4 / AR5	8 / 12 / 0.5	15 / 30 / 0.5	2.3	7
AR6 (Projection)	*5000/1000/5*	*???*	*???*	*250 Mbit/sec*

- Number of users increased in factor 5 for last 5 years
- Slowdown in proportion of data used to data saved probably can be expected

CMIP6: all tiers/priorities

- Experiments: 200

  5 times ↑
- CMIP Tables: 45

  2 times ↑
- Fields: 2000

  2 times ↑
- Years: 48000

  8 times ↑
- GFDL Total Amount: 1 PB

  5 times ↑
- Total Amount (all centers): 15 PB[1]

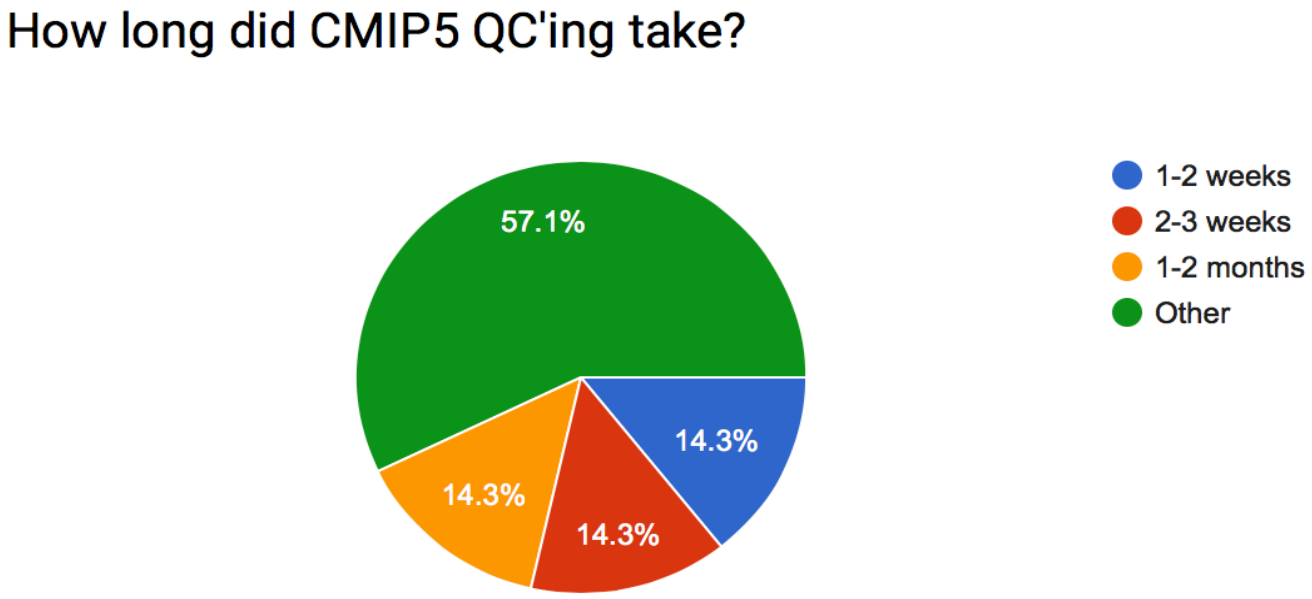
  9 times ↑

---

[1] According to WGCM-20 Questionnaire and calculated by Martin Juckes Python Library dreqPy

Human Costs: Data Producers Side

GFDL Poll:
CMIP5 QC Efforts and Suggestion



CMIP5 QC Human Costs:

- CMIP5 took from 2 weeks to 1 year of life of 10 scientists

GFDL Poll:
CMIP5 QC Efforts and Suggestion (cont.)

Efficient Tools Used:

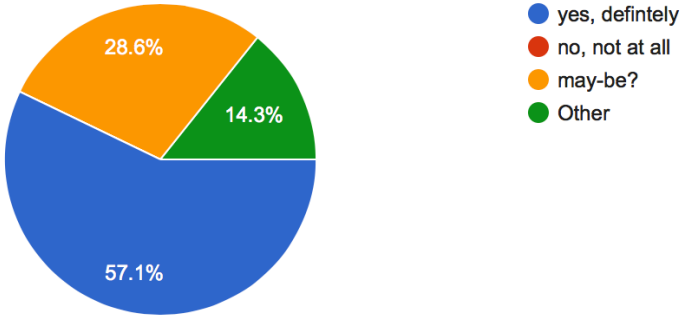
- Statistics: average, variance, global integrals.
- Number of missing values, valid range, ability to check orientation (N versus S, top vs. bottom).
- Many ferret scripts and statistics in Curator.
- Curator tools bookkeeping QC of big sets of files and integrated publishing automation.

Need to provide:

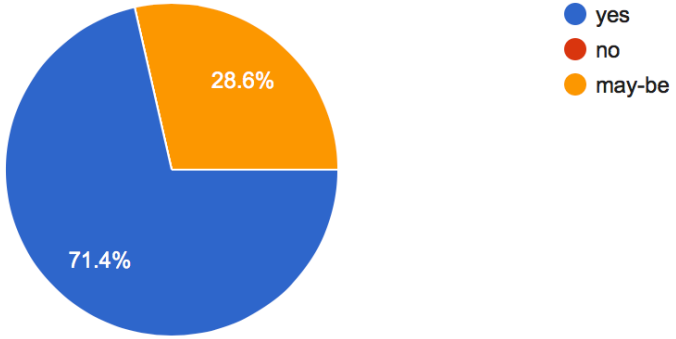
- ipython notebook
- Visual tools to help identifying outliers in variables

Scientific standards for QC

For data volumes that are projected to be higher than CMIP5 for CMIP6, do you think a agreeing upon "standardized QC checks/techniques" for users to adhere to might be more efficient?



Did Curator tools and MDBI help with some level of QC'ing?



Data Producers Costs: CMIP6 Expectation (GFDL)

Computational Resources

(courtesy to I.Held GFDL MDT presentation)

Earth System Model

- 15 SY/day, 5K cores
- 2KY DECK+ => 8% of resources for 4 months
- 10KY MIPS => 25% for 6 months

Higher resolution Physical Model

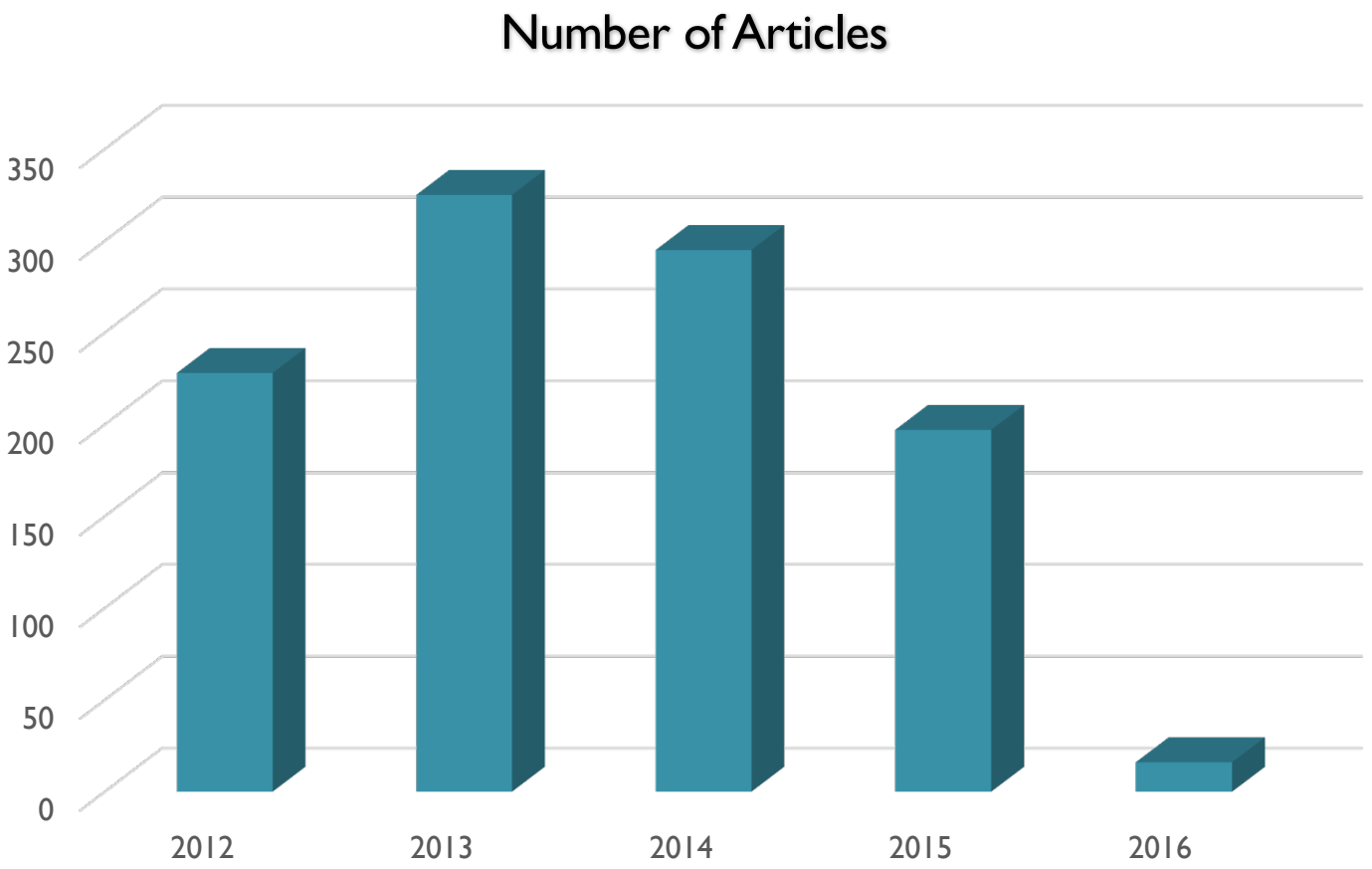
- 14 SY/day, 7.5K cores
- 2KY DECK+ => 8% for 4 months

Human Costs

- CMIP6 will require at least in 2 times more people
- More than half of scientists considers standard policy is definitely needed for scientific part of QC

Data Consumers Efforts for Extracting Science

CMIP5 Science Making Dynamics
(from *https://cmip-publications.llnl.gov*)



CMIP6 Science Projections

- **1000** scientists were participating in articles on IPCC AR5
- **2500** articles were written for this period.
- If all data was used then ~**700** TB per article were utilized.
- IPCC AR6 will have at least 10 times more data. Linear extrapolation gives abnormal number of articles - **25000** and **10000** scientists required (assuming that output resolution will be on a par with AR5). Obviously, it will not happen and either way – big part of data will not be claimed ever or each article will require more data & more time for data analysis.
- Rhetorical question: Is climate community capable to ingest such amount of data for 6 years?

Some Conclusions

- CMIP6 will be a serious challenge for IT maturity to serve such immense climate project.
- Ensure tight harmonized cooperation of data producers, data administrators (publishers) and data analyzers to make sure that goals are capable of being met by all parties.
- Need to standardize scientific part of QC policy for all modeling centers. It will increase data credibility.
- Automation of scientific QC is vital necessity.
- Needs to make variables tracking which were used. It will be good base for next CMIP planning.
- Regridding output data of all centers to uniform grid (the same type and resolution) will increase substantially data usability.